Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Human Consistency Evaluation of Static Video Summaries

Sivapriyaa Kannappan[a] ·
Yonghuai Liu[a,b] · **Bernard Tiddeman**[a]

Received: date / Accepted: date

Abstract Automatic video summarization aims to provide brief representation of videos. Its evaluation is quite challenging, usually relying on comparison with user summaries. This study views it in a different perspective in terms of verifying the consistency of user summaries, as the outcome of video summarization is usually judged based on them. We focus on human consistency evaluation of static video summaries in which the user summaries are evaluated among themselves using the consistency modelling method we proposed recently. The purpose of such consistency evaluation is to check whether the users agree among themselves. The evaluation is performed on different publicly available datasets. Another contribution lies in the creation of static video summaries from the available video skims of the SumMe datatset. The results show that the level of agreement varies significantly between the users for the selection of key frames, which denotes the hidden challenge in automatic video summary evaluation. Moreover, the maximum agreement level of the users for a certain dataset, may indicate the best performance that the automatic video summarization techniques can achieve using that dataset.

Keywords Video Summarization · Keyframe Extraction · Performance Evaluation · Consistency modelling · User Consistency

Sivapriyaa Kannappan[a]
E-mail: sik2@aber.ac.uk

Yonghuai Liu[a,b]
E-mail: yyl@aber.ac.uk

Bernard Tiddeman[a]
E-mail: bpt@aber.ac.uk
[a] *Department of Computer Science, Aberystwyth University, Ceredigion SY23 3DB, UK*
[b] *Department of Computer Science, Edge Hill University, St Helens Road, Ormskirk, Lancashire L39 4QP, UK*

1 Introduction

Due to the recent advancements in multimedia technologies and widespread use of internet amenities over recent years, the amount of video contents produced and their archives keep growing rapidly. Among the multimedia types (such as text, image, graphic, audio and video), video is the most challenging one as it combines all the other media data into a single data stream and it is not easy to gain efficient access due to its unstructured format and variable length [1]. This creates the need for succinct representation of videos in order to browse it more effectively and efficiently. The limited storage space available also demands video summarization without losing significant information. Video summarization aims to extract distinct key frames within a video, by providing a representative summary with reduced file size.

According to Troung and Venkatesh [2], there are two kinds of video summaries: *Static video summary*, which incorporates a set of key frames and *Dynamic video skimming*, which consists of a set of shots extracted from the original video [3]. The major benefit of video skimming is that the summary includes both audio and motion elements, which enrich the emotions and the amount of information delivered by the summary. It is also stimulating to view a skim with an audio-visual component instead of photographic slides of static key frames. In contrast, as static video summaries are not confined to timing and synchronization issues, it is more pliable compared to sequential display of video skims [3]. Moreover, static video summaries encompass both spatial and temporal information (key events in a specific order) which enables the user to rapidly grasp the video content, thereby reducing the computational complexity for various applications such as video retrieval and browsing, navigation and indexing [2]. In fact, the temporal order is usually not maintained while the static video summarization is generated using clustering techniques, but it can still be recovered by automatic ordering of extracted key frames based on the frame index of the video summary. Therefore, we concentrate on static video summaries.

Traditional static video summarization approaches lie under two major categories: (i) Shot-based, where video shots are detected initially and then key frames are extracted from each shot. Zhang *et al.* [4] proposed a shot detection method based on image histogram which has been extensively used for frame representation in the research community [5] [6]. (ii) Segment-based, where the video is sliced into segments (may be a scene or combination of one or more shots) and key frames are extracted from each segment. Yeung *et al.* [7] proposed a graphical representation of videos by constructing a *scene transition graph* where each node represents a shot and the edges represent the transitions between the shots based on the visual similarity and temporal locality [8] [9].

Various video summarization techniques are proposed in the literature, incorporating color, motion and textual features. Though, audio carries significant information in a video, it is not always beneficial to consider the acoustic features for video summarization [10]. While end-to-end trainable deep

Convolutional Neural Networks (CNN) based bi-directional Long-Short Term Memory (LSTM) model and multi-task learning [11,12] is proposed for image caption generation and image retrieval tasks, the extracted features may be used to represent the video frames for video summarization. In another study, LSTM is used as a frame selector for unsupervised video summarization in the generative adversarial networks (GAN) based on recurrent auto-encoders [13]. Moreover, deep CNN features [14,15], convolutional features integrated with Prewitt and Sobel edge detectors [16] and Haar-like features [17] may also be applied to represent video frames for effective video summarization.

However, video summary evaluation is challenging due to the lack of a common evaluation strategy. According to Troung and Venkatesh [2], the existing video summary evaluation techniques can be grouped into four main categories: (i) Result description [1], [18]: This is the most common form of evaluation as it does not involve any comparison with other techniques; (ii) Objective metrics [19]: Here the metric is often the fidelity function computed between the automatic summary and the original frame sequence for the measurement of the extent to which the former can approximate the latter. However, there is no experimental justification of how well the metric maps to human judgement regarding the quality of the automatic video summaries; (iii) User studies [20]: Here the evaluation employs user studies where independent users judge the quality of automatic video summaries. Even though this is the most useful and realistic form of evaluation, it is not commonly used due to the difficulty in setting them up; and (iv) User summaries [3]: A novel evaluation method called Comparison of User Summaries (CUS) is proposed in [3] where the video summary is built by a number of users from the sampled frames. Those user summaries serve as a ground truth for their comparison with the automatic summaries obtained by different approaches.

Though there are many possible ways to summarise a video, the most expected summary depends on the application and the desired length. Consequently, the evaluation of those video summaries are tricky and usually subjective. User summaries are widely used as a reference in [3],[21–24] for the calculation of precision, recall, f-measure, and dice coefficient, for example, to quantify the performance of the automatic techniques.

Among the video summary evaluation techniques, De Avila *et al.* [3] and Mei *et al.* [21] compute the match between automatic and a user summary using color features, Mahmoud [22] and Mahmoud *et al.* [23] integrate both color and texture features, Sharghi *et al.* [25] use intersection-over-union (IOU) similarities defined directly over the user annotated semantic vectors as edge weights in order to find the semantic match. Kannappan *et al.* [26] incorporate the Efficient Image Euclidean Distance (EIMED) for the search of matched frames between the automatic summary and a user summary. We intend to use our own evaluation method [27] due to its simplicity, reliability and efficiency in identifying the more faithful matched frames through compatibility modelling (which incorporates correlation and consistency estimation).

Though several evaluation strategies exist, our interest in this study is to investigate the consistency of user summaries among different datasets and

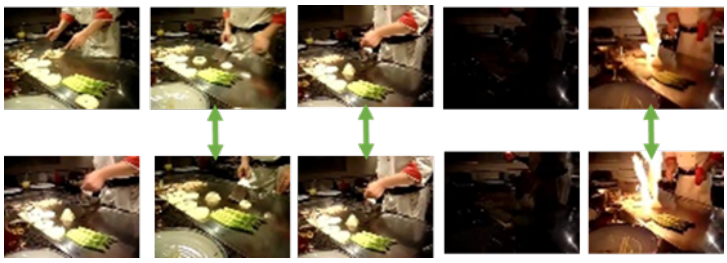

Fig. 1 Potential matches between User summary #4 *(top)* and the Overall trimmed Summary *(bottom)* of the video *Cooking* in the SumMe dataset. Arrows represent the refined matches after compatibility modelling.

to evaluate the extent to which the users agree among themselves. It is not the intention of the study to propose new methods to create a summary from a given video, but to investigate how one summary can be better evaluated against another. To this end, firstly we evaluate the user summaries using the evaluation method proposed by Kannappan *et al.* [27]. The method is composed of two main steps: In the first step, initial matches are obtained between all the frames selected from the two different user summaries via a two-way search using correlation coefficient for the measurement of their similarity. Subsequently, consistency check is carried out where the inter-frame difference between different user summaries are maintained, leading to the identification and elimination of weak and false matches. To further reveal the consistency between the user summaries, we also calculate as an indicator the pairwise correlation between the number of frames chosen by different users.

In Fig. 1, 5 pairs of frames were computationally matched (using the first step in our evaluation method) between user #4 and the overall trimmed summary in the SumMe dataset. Though the first and the fourth pair of potential matched frames seem to be semantically similar, they are not pertinent matches as the frames differ in the orientation of the chef while cooking and a change in the brightness of a scene respectively. Such false matches will clearly distort and mislead the performance measurement, thus should be discarded. Most of the existing summarization techniques evaluate using timestamps of the key frames, in order to handle the inconsistent ground truth annotations, however it will not be appropriate for certain videos with more frequent shots. In this case, how to define the matches between the user summaries and automatically selected frames clearly plays a crucial role in revealing their true performance. The term 'match' in this paper denotes two images/frames that are the same. Those correct matches should maintain the difference between the potential matched pairs of frames (obtained using correlation) in the individual user summaries as shown in Fig. 1.

To validate the consistency among the users, three publicly accessible datasets are used. The experimental results do show that the level of agreement between different users varies drastically and such maximum agreement

level probably indicates the best performance that the automatic video summarization techniques can achieve.

The main contributions of this study can be summarised as:

1. An investigation of human consistency to reveal the hidden challenge within the video summary evaluation: (i) Consistency evaluation based on the *potential matched frames* between different user summaries; (ii) Consistency evaluation based on the *refined matched frames* using consistency modelling; and (iii) Consistency evaluation based on the correlation between the *numbers of frames* chosen by different users;
2. Static key frame creation from the annotated video segments of the SumMe dataset (SM) in order to conduct experiments over them along with the other two publicly available datasets.

The rest of this paper is structured as follows: Section 2 details our human consistency evaluation approach; Section 3 describes the performance metrics used; Section 4 explains the method of key frame creation for the SM dataset from the available video skims; Section 5 presents the experimental evaluation using three publicly available datasets; and finally Section 6 draws conclusions and indicates future work.

2 The Human Consistency Evaluation Method

Due to the lack of universal evaluation strategy for video summarization, the evaluation is indeed demanding. Since the outcome of the evaluation is usually judged based on the user summaries, we thus examine in this paper the consistency of those user summaries, so that it can reveal to what extent the current evaluation methods are reliable.

For a video v_i $(i = 1, 2, \cdots, n)$, N users were invited to manually select the key frames $u_i^j = \{u_i^{jj'} | j' = 1, 2, \cdots, a_{ij}\}$, where a_{ij} is the number of frames selected by user j $(j = 1, 2, \cdots, N)$ over v_i. A user summary u_i^j has been evaluated pairwise with all the other user summaries u_i^k over all the datasets used, in order to (i) retrieve the initial matches, (ii) refine the matches and (iii) find the correlation between the number of frames chosen. These steps are detailed in the following subsections.

2.1 Identification of potential matched frames between different user summaries using correlation

Here, the potential matched frames are obtained using a two-way search from a user summary u_i^j to a user summary u_i^k and then back to user summary u_i^j again. The two-way search is implemented as follows: We use one frame u_i^{jp} from the user summary u_i^j to search through another user summary u_i^k for the most similar frame $u_i^{kq'}$ and then use this most similar frame $u_i^{kq'}$ to search through the user summary u_i^j for the most similar frame $u_i^{jp'}$. If this

most similar frame $u_i^{jp'}$ is the same as the original frame u_i^{jp}, then we consider them as a match $(u_i^{jp}, u_i^{kq'})$. This process continues for all the frames within the user summary u_i^j, leading all the matched frames $u_i(j,k)$ to be identified: $u_i(j,k) = \{(u_i^{jp}, u_i^{kq'})\}$. The similarity $C1_i(p,q)$ between two frames u_i^{jp} and u_i^{kq} is measured using the Pearson's correlation coefficient as given in Equation 1:

$$C1_i(p,q) = \frac{\sum_{r=1}^{n}(x_r^p - \bar{x^p})(y_r^q - \bar{y^q})}{\sqrt{\sum_{r=1}^{n}(x_r^p - \bar{x^p})^2}\sqrt{\sum_{r=1}^{n}(y_r^q - \bar{y^q})^2}} \tag{1}$$

where x_r^p, y_r^q are the intensity values of the r^{th} pixel of u_i^{jp} and u_i^{kq} respectively, $\bar{x^p}$ and $\bar{y^q}$ are their mean intensity values. The benefit of the correlation coefficient is that it reduces the comparison of two two-dimensional images to a single scalar value between $[-1, 1]$ [28]. A negative value of the coefficient shows a negative correlation between the two images, a zero value shows no correlation, and a positive value shows a positive correlation. The larger the coefficient, the higher the similarity of the two images.

2.2 Refinement of potential matched frames using the consistency modelling

The matched frames obtained in the last section are not always reliable and thus usually need to be refined. Hence, we believe that all the correct matches should maintain the difference between the matched pairs of frames in the individual user summaries. In contrast, such difference may not be maintained for the weak and false matches.

To this end, a match $(f(m), f'(m))(m = 1, 2, \cdots, |u(j,k)|)$ between two user summaries (u^j, u^k) over a particular video is represented using two three-dimensional histograms h_m and h'_m in the HSV color space for each matched frame from each user summary, where 16, 4 and 4 bins were used for Hue, Saturation, and Value respectively. Initially, we calculate the difference d_{ml} and d'_{ml} between any two matched frames $(f(m), f'(m))$ and $(f(l), f'(l))$ between any two user summaries: $d_{ml} = ||h_m - h_l||$ and $d'_{ml} = ||h'_m - h'_l||$. Then the compatibility matrix $C = \{C_{ml}\}$ is calculated as follows:

$$C_{ml} = \exp\left(-s * |d_{ml} - d'_{ml}|\right) >= 0 \tag{2}$$

where s is a stretch factor which controls how heavily the difference $|d_{ml} - d'_{ml}|$ should be penalised and it was set to 25 in this paper based on the experimental tests. The eigenvector x^* of C corresponding to the maximum eigenvalue [29] indicates the extent to which the frames are consistently matched.

$$x^* = \underset{x}{argmax}\frac{x^T C x}{x^T x} \tag{3}$$

It can be computed effectively using the iterative power method [30] as follows: An initial vector $x^{(0)}$ is randomly generated over the interval [-1, 1] and then multiplied by the compatibility matrix C to form another new vector and the process continues until the difference between $x_i^{(t)}$ and $x_i^{(t+1)}$ at two consecutive iterations t and $t+1$ is below a threshold of 0.00001, for example:

$$x^{(t+1)} = \frac{Cx^{(t)}}{\|Cx^{(t)}\|} \tag{4}$$

where $\|Cx^{(k)}\|$ denotes the Euclidean length of a vector. Finally, the relative consistency value $\hat{x_i}$ of the match $(f(m), f'(m))$ is calculated from $x^* = (x_1, x_2, \cdots, x_{|u(j,k)|})$ as: $\hat{x_m} \leftarrow x_m / x_{max}$ where $x_{max} = \max_{m'} x_{m'}$. If the relative consistency value $\hat{x_i}$ is below a threshold δ, 0.6 for example, then the match $(f(m), f'(m))$ should not be treated as a correct one for the performance measurement of those two user summaries u^j and u^k. It should be noted that the stretch factor s and the threshold δ may be data dependent and the setup of these parameters are detailed in [27].

2.3 Consistency evaluation based on the number of frames chosen by different users using correlation

Different users usually select different numbers of frames from a given video for their summary. In this case, the consistency between two users may be measured using the correlation between the number of frames selected. This measure has an advantage of easy implementation that does not involve the difficult task of establishing frame matches. The correlation between the numbers m_{ij} and m_{ik} of selected frames by two users j and k over different videos v_i is computed as:

$$C2(j,k) = \frac{\sum_{i=1}^{n}(a_{ij} - \bar{a_j})(a_{ik} - \bar{a_k})}{\sqrt{\sum_{i=1}^{n}(a_{ij} - \bar{a_j})^2}\sqrt{\sum_{i=1}^{n}(a_{ik} - \bar{a_k})^2}} \tag{5}$$

where $\bar{a_j} = \frac{1}{n}\sum_i a_{ij}$ and $\bar{a_k} = \frac{1}{n}\sum_i a_{ik}$ are the mean number of key frames selected by user j and k over all the videos respectively.

3 Performance Metrics

We adopt the widely used metrics such as precision, recall and f-measure [31] for the evaluation of the performance of different users using the initial and the refined matches obtained from Section 2, where precision determines how many selected frames are relevant, whereas recall determines how many relevant frames are selected. F-measure is the harmonic mean of both precision

and recall, which reveals the accuracy of a user summary. The higher the F-measure, the better the accuracy.

$$p_i(j,k) = \frac{|u_i(j,k)|}{a_{ij}}, \tag{6}$$

$$r_i(j,k) = \frac{|u_i(j,k)|}{a_{ik}}, \tag{7}$$

$$f_i(j,k) = 2 * \frac{p_i(j,k) * r_i(j,k)}{p_i(j,k) + r_i(j,k)}. \tag{8}$$

where $p_i(j,k)$, $r_i(j,k)$ and $f_i(j,k)$ are precision, recall and f-measure between a user summary u_i^j and another user summary u_i^k over video v_i respectively. $|u_i(j,k)|$ is the number of matched frames between u_i^j and u_i^k, a_{ij} and a_{ik} are the number of frames selected by user j and k over video v_i respectively. The consistency of user summaries are evaluated using the pairwise f-measure between them as performed by [24]. The pairwise mean precision $p(j,k)$, recall $r(j,k)$ and f-measure $f(j,k)$ between two user summaries u_i^j and u_i^k over all the videos v_i are defined as:

$$p(j,k) = \frac{1}{n}\sum_{i=1}^{n} p_i(j,k), r(j,k) = \frac{1}{n}\sum_{i=1}^{n} r_i(j,k),$$
$$f(j,k) = \frac{1}{n}\sum_{i=1}^{n} f_i(j,k). \tag{9}$$

The mean precision $p(j)$, recall $r(j)$, and f-measure $f(j)$ of a particular user j against all the other users k are computed as:

$$p(j) = \frac{1}{N-1}\sum_{k=1,k\neq j}^{N} p(j,k), r(j) = \frac{1}{N-1}\sum_{k=1,k\neq j}^{N} r(j,k),$$
$$f(j) = \frac{1}{N-1}\sum_{k=1,k\neq j}^{N} f(j,k). \tag{10}$$

From the metrics above, it can be clearly seen that the method used to find the matched frames $u_i(j,k)$ plays a crucial role in the measurement of the performance of different techniques. Thus, it must be accurately defined. The incorrect matches will lead to inaccurate and misleading performance measurement for different techniques.

4 Keyframe creation from the SM datatset

Initially, key frames are extracted for each individual user summary (from the available video skims summarized by 15 to 18 different people, from which we chose the beginning 15 users to make it consistent for all the 25 videos in the SM dataset [24]), by sampling the middle frame from each skim excerpt of a

Table 1 Extracted key frames from the video skims chosen by User #1 for the videos *Eiffel Tower.mp4* and *Jumps.mp4* from the SM dataset.

User #	Video Name	Video Skims	Extracted Key Frames
1	Eiffel Tower	1190-1394	1292,
		2458-2606	2532,
		3945-4130	4038.
1	Jumps	49-85	67,
		306-361	334,
		473-526	500.

video [2]. Here, we considered only the beginning 15 user annotations for each video in order to ensure consistent evaluation. Subsequently, the overall user summary (US) is created based on the votes of the users for each frame in a video. Considering a video, an array of per frame votes is searched to find regions of maxima with a magnitude greater than or equal to 8 (more than 50 %). The first frame of each such stretch is chosen as the key frame (in overall US) for each varying user votes. In order to avoid neighboring frames being selected, if the difference in index between the consecutive frames in the overall user summary is less than 25, then those frames are not considered as key frames, thereby are eliminated. Such a process leads to a non-redundant overall trimmed summary (TS). Examples of the key frame creation from the video skim (by user #1) excerpt and overall US & TS key frame creation for the videos *Eiffel Tower* and *Jumps* are shown in Tables 1 and 2 respectively. The number of key frames extracted from the video segment chosen by various users, overall US and their TS for all the 25 videos are presented in Table 3.

Table 2 Overall user summary (US) and Overall trimmed summary (TS) key frame creation for the videos *Eiffel Tower* and *Jumps* from the SM dataset.

Video Name	Frame # of Overall US	# of Users Extracted The Frame	Final Key Frame Selected (Overall TS)
Eiffel Tower	118	8	✓
	126	9	
	127	10	
	150	9	✓
	154	8	
	1216	9	✓
	1224	10	
	1233	11	
	1238	12	
	1242	13	✓
	1261	12	
	1268	10	✓
	1309	9	✓
	1325	10	
	1333	9	
	1348	10	✓
	1349	9	
	1365	8	
	4884	9	✓
	4900	8	
Jumps	61	8	✓
	63	9	
	64	10	
	72	9	
	74	8	
	304	10	✓
	306	11	
	308	12	
	317	13	
	318	14	
	328	13	
	335	14	✓
	360	13	✓
	361	12	
	362	10	
	364	9	
	366	8	

Table 3 Total # of key frames extracted from each video chosen by different Users, overall user summary (US) and overall trimmed summary (TS) of the SM dataset.

Video Name	Total # of frames	U #1	U #2	U #3	U #4	U #5	U #6	U #7	U #8	U #9	U #10	U #11	U #12	U #13	U #14	U #15	Overall US	Overall TS
Air force one	4494	4	4	7	11	3	8	3	4	2	10	3	5	3	10	3	16	9
Base jumping	4729	5	9	2	6	2	5	7	7	6	4	5	5	6	3	6	12	5
Bear park climbing	3341	6	4	7	4	8	2	3	7	3	3	8	6	2	7	6	2	1
Bike polo	3064	1	5	1	4	5	4	4	3	3	4	4	5	5	4	6	17	6
Bus in rock tunnel	5133	5	3	11	3	8	4	10	7	7	6	4	5	3	5	2	7	3
Car over camera	4382	2	6	9	7	4	2	5	8	3	1	4	3	2	2	9	8	5
Car rail crossing	5075	4	10	5	3	4	8	4	5	4	7	5	4	3	5	2	25	11
Cockpit landing	9046	8	6	11	10	5	2	9	10	5	6	9	1	7	5	9	14	8
Cooking	1287	3	3	3	6	3	3	4	3	4	3	1	3	4	2	3	27	7
Eiffel tower	4971	3	7	7	6	5	7	6	5	5	6	3	4	12	3	4	20	8
Excavators river crossing	9721	9	6	18	6	5	13	17	15	6	1	6	10	9	14	7	27	13
Fire domino	1612	2	5	5	2	3	3	9	3	3	4	3	4	8	3	3	21	5
Jumps	950	3	3	5	4	1	2	3	1	3	3	3	4	2	4	3	17	4
Kids playing in leaves	3187	4	1	3	5	7	4	2	3	4	3	4	8	1	10	3	10	5
Notre dame	4608	10	7	14	5	3	8	6	8	3	11	4	8	16	7	5	13	5
Paintball	6096	4	10	6	9	6	3	5	4	3	7	4	3	4	3	3	25	10
Paluma jump	2574	3	4	4	3	3	5	3	1	3	4	1	3	2	3	3	28	8
Playing ball	3119	6	6	10	7	1	2	6	3	6	4	3	2	6	3	5	12	6
Playing on water slide	3065	6	4	2	9	2	5	6	3	3	5	13	3	3	10	3	1	1
Saving dolphins	6683	7	4	3	8	5	9	5	6	8	10	9	1	10	5	7	9	5
Scuba	2221	5	3	3	5	2	3	3	5	1	3	3	4	4	3	6	8	5
St. Maarten landing	1751	2	4	4	3	2	2	4	3	1	1	4	1	7	3	4	19	7
Statue of liberty	3863	2	5	3	2	2	2	1	11	3	2	3	5	5	2	1	5	4
Uncut evening flight	9672	6	8	7	3	7	8	3	4	3	13	3	6	4	6	9	28	11
Valparaiso downhill	5178	10	10	16	2	10	3	7	10	7	2	5	10	13	5	6	20	9

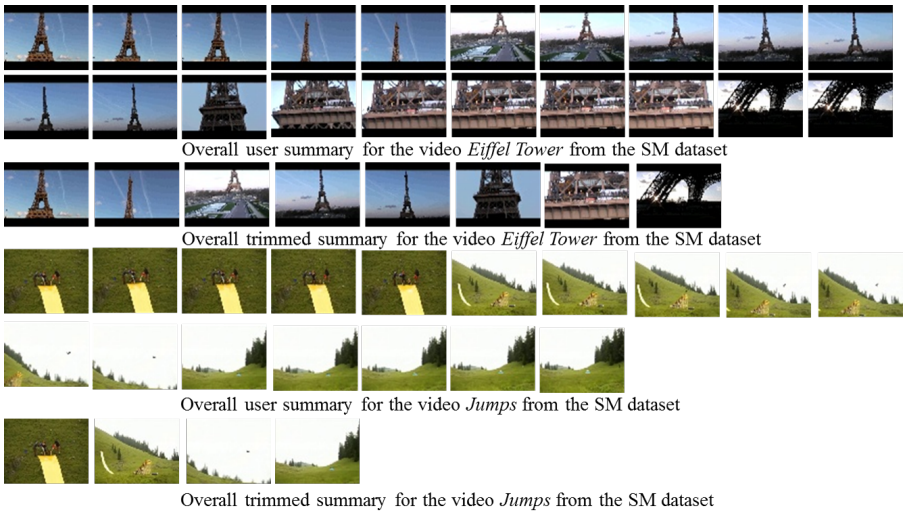

Fig. 2 Key frames created for the overall user summary and overall trimmed summary for the videos *Eiffel Tower* and *Jumps* in the SM dataset.

Fig. 2 shows the selected key frames from the videos *Eiffel Tower* and *Jumps*, including the overall US and the overall TS. The first 2 rows with 20 frames depict the overall US for the video *Eiffel Tower* and the 3^{rd} row with 8 key frames depicts the overall TS. Also the 4^{th} and 5^{th} rows with 17 frames depict the overall US for the video *Jumps* and the 6^{th} row with 4 key frames depicts the overall TS. It can be seen that the similar frames are eliminated based on the proposed process.

5 Experimental Results

In this section, we evaluate the consistency of user summaries using three publicly available datasets captured by either professionals or amateurs. All the experiments were carried out on a Intel core i7, 3.60 GHz computer with 8GB RAM.

5.1 Three Datasets Used

The first dataset was 50 videos selected from the Open Video (OV) Project [32]. The selected videos are in MPEG-1 format containing 30 fps with a resolution of 352×240 pixels. The videos incorporate several genres (documentary, ephemeral, historical, lecture) and their duration varies from 1 to 4 minutes. Those videos were also used by [33, 20, 34, 3].
The second dataset was 50 videos from the YouTube (YT) database, which differ in color, length, motion and theme (eg., cartoons, news, sports, commercials, tv-shows and home videos) created by [3] and their duration varies

Table 4 Mean F-measure of the user summaries based on the potential matched frames for 50 videos from the **OV dataset**.

	Mean F-measure				
User #	1	2	3	4	5
1	1	0.74	0.74	0.75	0.74
2	0.74	1	0.78	0.75	0.75
3	0.74	0.78	1	0.77	0.76
4	0.75	0.75	0.77	1	0.72
5	0.74	0.75	0.76	0.72	1

Table 5 Mean F-measure of the user summaries based on the potential matched frames for 50 videos from the **YT dataset**.

	Mean F-measure				
User #	1	2	3	4	5
1	1	0.68	0.65	0.64	0.69
2	0.68	1	0.67	0.66	0.69
3	0.65	0.67	1	0.66	0.67
4	0.64	0.66	0.66	1	0.66
5	0.69	0.69	0.67	0.66	1

from 1 to 10 minutes. The user study conducted by De Avila et al. [3] for both the OV and YT video datasets were used as ground truth summaries, where the user summaries were created by 50 users, each one dealing with 5 videos, meaning that each video has 5 different user summaries, so in total 250 summaries were created manually [3].
Finally, the third dataset was 25 videos chosen from the SM dataset [24]. They are raw or minimally edited user videos (comprising holidays, events and sports) and their duration varies from 1 to 6 minutes. The user study conducted by Gygli *et al.* [24] over the SM dataset were used to extract key frames from the video segments in which each video was summarized by 15 to 18 different people. However, we considered only the beginning 15 user annotations for each video in order to ensure consistent evaluation.

5.2 Consistency evaluation based on the potential matched frames

The pairwise mean f-measure in Equation 9 over the potential matched frames between different user summaries lie between **0.72** and **0.78** for the OV dataset, **0.64** and **0.69** for the YT dataset and **0.43** and **0.50** for the SM dataset which can be seen from Tables 4 and 5 and Fig. 5 respectively. The overall means of mean f-measure are **0.75, 0.67 and 0.46** as shown in Figs. 3, 4 and 5 for the OV, YT and SM dataset respectively. Fig. 6 shows 5 different user summaries of the video *Hurricane Force - A coastal perspective, segment 3* in the OV dataset. Apparently, only the first and the last frame in user #5 and the 4^{th} frame in user #1 are different. However, a careful observation of the 7^{th} frame in user #1 and user #2 respectively shows that there is also a difference in the landscape but, it is considered as a match by the two-way

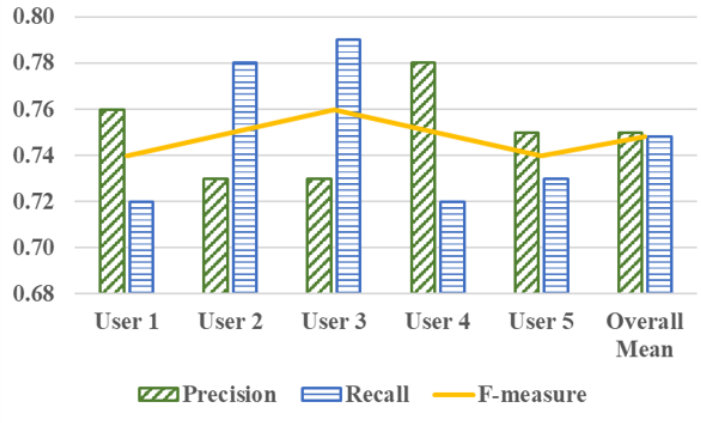

Fig. 3 Overall performance metrics of the user summaries based on the potential matched frames for 50 videos from the **OV dataset**.

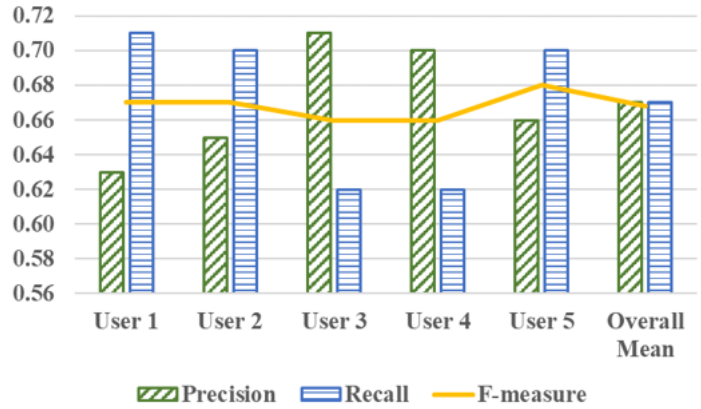

Fig. 4 Overall performance metrics of the user summaries based on the potential matched frames for 50 videos from the **YT dataset**.

search. Similarly, among the corresponding 3^{rd} frames selected by user #1 and user #2, the frame in user #1 is slightly different in the rectangular display along with some background, which is again considered as a match by the two-way search. These weak and false matches clearly lead to an optimistic performance measurement of different users and thus should be eliminated in order to obtain a more accurate and objective performance measurement, which will be investigated below in Section 5.3.

Fig. 7 shows 5 different user summaries of a cartoon video in the YT dataset. It contains 9, 16, 17, 10 and 16 frames for user #1, #2, #3, #4 and #5 respectively, which show the varying level of agreement between users even among the number of frames chosen. Also, if we consider the actual user summaries extracted by different users, they differ drastically. Suppose, if we

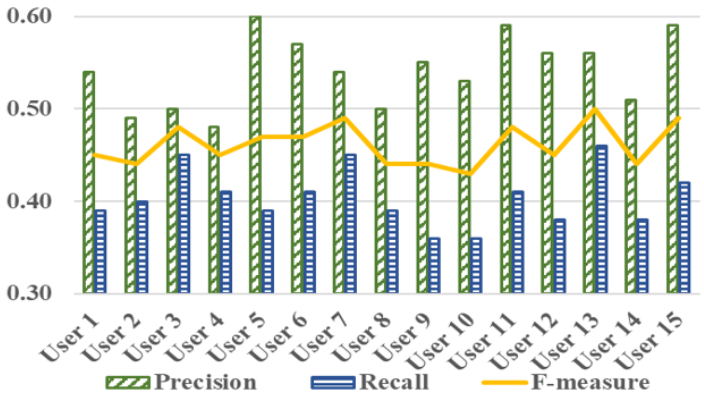

Fig. 5 Overall performance measures of the user summaries based on the potential matched frames for 25 videos from the **SM dataset**.

Fig. 6 User summaries of the video *Hurricane Force - A coastal perspective, segment 3* in the **OV dataset**.

consider user #1 summary (containing 9 frames), the 1^{st}, 2^{nd}, 3^{rd}, 5^{th}, 6^{th} and 8^{th} frames occur in all the other user summaries, whereas the 4^{th} frame is missing in those of user #3, #4, and #5 and 7^{th} and 9^{th} frames are missing in those of user #4. Also a varying large number of frames are extracted by users #2, #3 and #5 respectively. Fig. 8 shows 15 different user summaries of the video *Cooking* in the SM dataset. Fig. 9 contains 6 frames as user summary #4 (at the top) and 7 frames as overall trimmed summary (at the bottom), in which the arrows show the corresponding 5 matches between them. While considering both the number of frames and the actual user summaries extracted by different users, only half of them agree among themselves which reveals that the users' agreement is better in the OV dataset compared to the

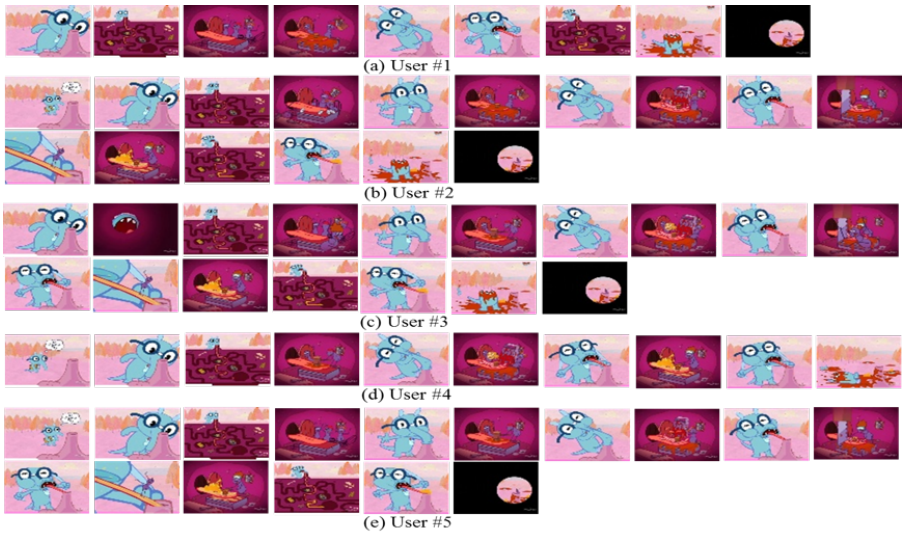

Fig. 7 User summaries of a cartoon video in the YT dataset.

Table 6 Mean F-measure of the user summaries based on the refined matched frames for 50 videos from the **OV dataset**.

	Mean F-measure				
User #	1	2	3	4	5
1	1	0.62	0.63	0.65	0.64
2	0.63	1	0.69	0.66	0.68
3	0.64	0.69	1	0.67	0.68
4	0.65	0.67	0.68	1	0.65
5	0.65	0.67	0.69	0.65	1

YT and SM datasets. This may be due to the professionalism and organization in capturing the videos of the OV dataset, which leads various users to select almost similar key frames. For the SM dataset, the correlation goes even to negative (as shown in Fig. 14), because of the diverse content and amateur quality of the videos.

5.3 Consistency evaluation based on the refined matched frames

The pairwise mean f-measure in Equation 9 based on the refined matched frames lie between **0.62** and **0.69** for the OV dataset, **0.53** and **0.60** for the YT dataset and **0.39** and **0.46** for the SM dataset which can be seen from Tables 6 and 7 and Fig. 13 respectively. The overall means of mean f-measure are **0.66, 0.56 and 0.43** as shown in Figs. 11, 12 and 13 for the OV, YT and SM datasets respectively. It can be seen here that the performance measures tend to decrease compared to those in Tables 4 and 5 and Fig. 5. This is because they retain only the consistent strong matches between the

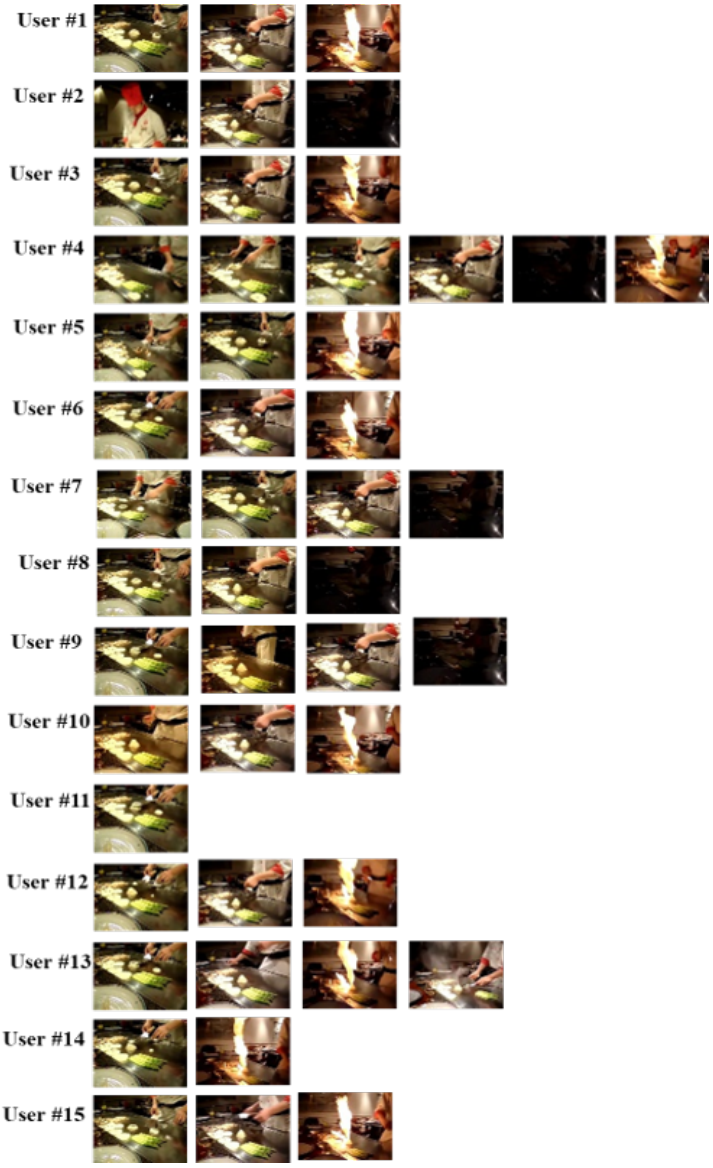

Fig. 8 User summaries of the video *Cooking* in the SM dataset.

pairwise user summaries as shown in Fig. 10. The normalised eigenvalues for the potential matches in Fig. 9 based on the normalized correlation coefficient are **0.543164**, 0.710773, 0.761488, **0.0533676**, 1 respectively. The first and the fourth eigenvalue (highlighted with bold font) is relatively small and indicates that it is a weak match. Visually, the first pair of matched frames represents a chef cooking in the kitchen, however the positions of the chef differ between

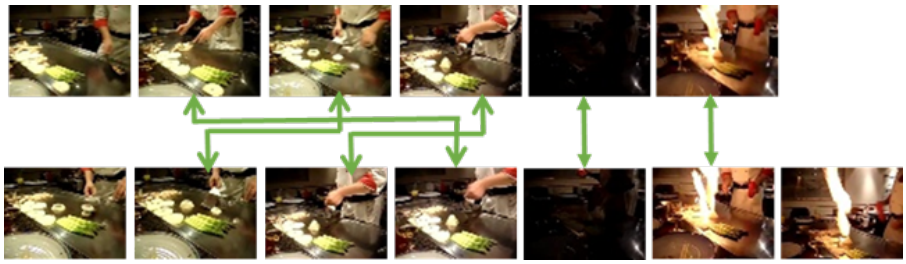

Fig. 9 User summary #4 *(top)* Vs Overall trimmed Summary *(bottom)* of the video *Cookings* in the SM dataset. Arrows represent the corresponding potential matches between them.

Fig. 10 Consistent matches between User Summary #4 *(top)* Vs Overall trimmed Summary *(bottom)* of the video *Cooking* in the SM dataset.

Table 7 Mean F-measure of the user summaries based on the refined matched frames for 50 videos from the **YT dataset**.

	Mean F-measure				
User #	1	2	3	4	5
1	1	0.58	0.54	0.55	0.60
2	0.58	1	0.56	0.57	0.57
3	0.54	0.55	1	0.53	0.55
4	0.55	0.58	0.53	1	0.56
5	0.60	0.59	0.56	0.57	1

the frames. Similarly, the fourth pair of matched frames represents the change from darkness to brightness, while the chef was lighting the flame. Thus, it is reasonable to eliminate both the matches.

5.4 Consistency evaluation based on the number of frames chosen by different users using correlation

The pairwise correlation in Equation 5 between the numbers of key frames chosen by different users is somewhat acceptable for the OV dataset, as their correlation values lie between **0.51** and **0.80** as shown in Table 8 and can also be observed in Fig. 6, where it contains similar numbers of frames. However, the consistencies of the YT and SM dataset are much worse as their correlation values lie between **0.10** and **0.85** for the YT dataset and **-0.21** and **0.20** for the

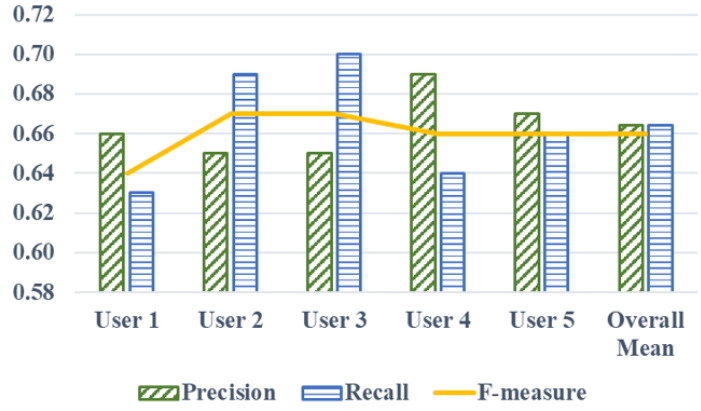

Fig. 11 Overall performance metrics of the user summaries based on the refined matched frames for 50 videos from the **OV dataset**.

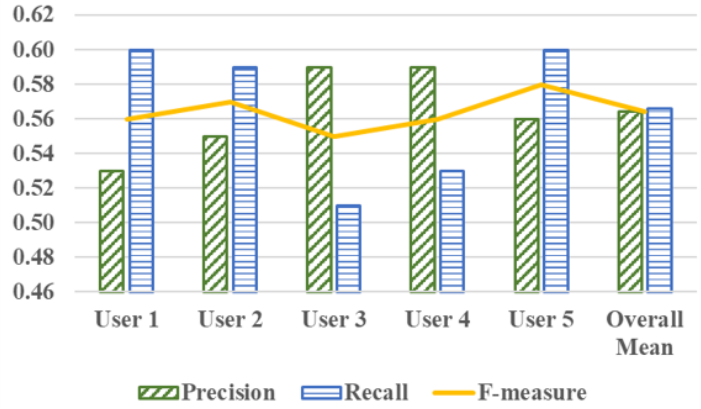

Fig. 12 Overall performance metrics of the user summaries based on the refined matched frames for 50 videos from the **YT dataset**.

Table 8 Correlation between the # of key frames chosen by different users for 50 videos from the **OV dataset**. Bold and Underlined values indicate minimum and maximum values respectively.

	Correlation				
User #	1	2	3	4	5
1	1	0.64	0.57	0.71	**0.51**
2	0.64	1	0.77	0.78	<u>0.80</u>
3	0.57	0.71	1	0.64	0.73
4	0.71	0.78	0.64	1	0.62
5	**0.51**	<u>0.80</u>	0.73	0.62	1

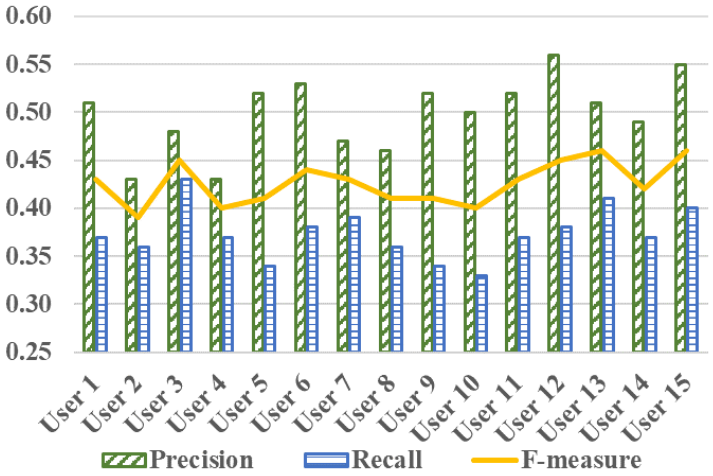

Fig. 13 Overall performance metrics of the user summaries based on the refined matched frames for 25 videos from the **SM dataset**.

Table 9 Correlation between the # of key frames chosen by different users for 50 videos from the **YT dataset**. Bold and Underlined values indicate minimum and maximum values respectively.

	Correlation				
User #	1	2	3	4	5
1	1	0.29	0.70	<u>0.85</u>	0.72
2	0.29	1	0.39	**0.10**	0.21
3	0.70	0.39	1	0.71	0.47
4	<u>0.85</u>	**0.10**	0.71	1	0.66
5	0.72	0.21	0.47	0.66	1

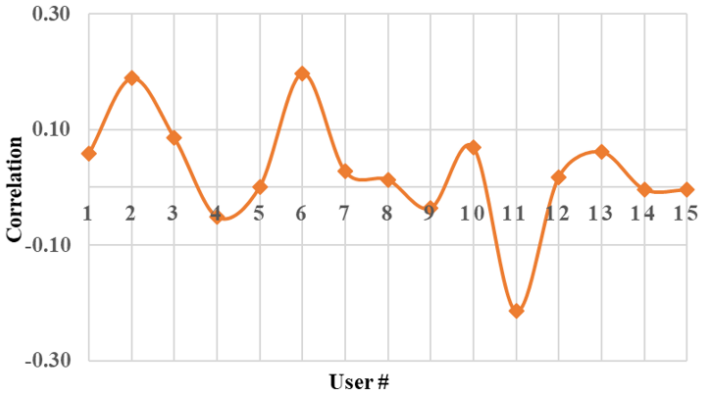

Fig. 14 Correlation between the # of key frames chosen by different users with the overall user summaries for 25 videos from the **SM dataset**.

Table 10 Execution time t in seconds taken for evaluation of the user summaries based on the refined matched frames for 50 videos from the **OV dataset**.

	Execution time *(t in sec)*				
User #	1	2	3	4	5
1	1	61	54	49	50
2	53	1	60	56	55
3	53	60	1	51	52
4	45	52	51	1	45
5	45	52	52	45	1

SM datset which can be seen from Table 9 and Fig. 14 respectively, where the bold and underlined values in the table indicate the minimum and maximum agreement level between the users. These results of correlation performed on the number of frames chosen by different users are consistent with those as demonstrated in Section 5.2. It reveals that the user summaries of the OV dataset are more consistent than those of the YT and SM datasets due to their well structuredness and similar contents.

5.5 Different similarity measures

In this section, we investigate the impact of different similarity measures on the establishment of the potential matched frames between different user summaries. To this end, the cosine similarity measure was also considered. We tried the cosine distance measure instead of the Pearson's correlation for finding the matches between the summary of User #1 and the overall trimmed summary for all the 25 videos in the SumMe dataset. The experimental results are presented in Figures 15 and 16.

Figure 15 shows that Pearson's correlation achieves a higher mean f-measure of 0.45 than that of 0.43 by the cosine distance. This is confirmed by Figure 16 where while the former established 2 matches between the summary of v5 by User #1 and the overall trimmed summary, the latter established only one. These results justifies our decision of using the Pearson's correlation for the establishment of the potential matched frames between different user summaries.

5.6 Computational Complexity

This work is devoted to the study of consistency of annotations, which is not required for the day-to-day use of summarization, hence the run time is not an issue. However, the establishment of the potential matches between two user summaries (u_i^j, u_i^k) over a particular video v_i through a two-way search have a complexity of $O(m_{ij}^2 m_{ik})$. The refinement of the potential matched frames has a computational complexity of $O(|u_i(j,k)|^2)$ where $|u_i(j,k)| \leq \min(m_{ij}, m_{ik})$. Hence, the overall complexity of the proposed human consistency evaluation over n videos is $O(nm^3)$. The evaluation is based on only a small set of key

Pearson's Correlation

Video #	KF	GT	Match	Precision	Recall	F-measure
1	4	9	2	0.5	0.222222	0.307692
2	5	5	2	0.4	0.4	0.4
3	6	1	1	0.166667	1	0.285714
4	1	6	1	1	0.166667	0.285714
5	5	3	2	0.4	0.666667	0.5
6	2	5	1	0.5	0.2	0.285714
7	4	11	2	0.5	0.181818	0.266667
8	8	8	3	0.375	0.375	0.375
9	3	7	3	1	0.428571	0.6
10	3	8	2	0.666667	0.25	0.363636
11	9	13	7	0.777778	0.538462	0.636364
12	2	5	2	1	0.4	0.571429
13	3	4	2	0.666667	0.5	0.571429
14	4	5	1	0.25	0.2	0.222222
15	10	5	2	0.2	0.4	0.266667
16	4	10	2	0.5	0.2	0.285714
17	3	8	2	0.666667	0.25	0.363636
18	6	12	3	0.5	0.25	0.333333
19	6	1	1	0.166667	1	0.285714
20	7	5	2	0.285714	0.4	0.333333
21	5	5	3	0.6	0.6	0.6
22	2	7	2	1	0.285714	0.444444
23	2	4	1	0.5	0.25	0.333333
24	6	11	3	0.5	0.272727	0.352941
25	10	9	3	0.3	0.333333	0.315789
Mean				**0.54**	**0.39**	**0.45**

Cosine Similarity

Video #	KF	GT	Match	Precision	Recall	F-measure
1	4	9	2	0.5	0.222222	0.307692
2	5	5	3	0.6	0.6	0.6
3	6	1	1	0.166667	1	0.285714
4	1	6	1	1	0.166667	0.285714
5	5	3	1	0.2	0.333333	0.25
6	2	5	1	0.5	0.2	0.285714
7	4	11	2	0.5	0.181818	0.266667
8	8	8	4	0.5	0.5	0.5
9	3	7	3	1	0.428571	0.6
10	3	8	1	0.333333	0.125	0.181818
11	9	13	6	0.666667	0.461538	0.545455
12	2	5	2	1	0.4	0.571429
13	3	4	2	0.666667	0.5	0.571429
14	4	5	1	0.25	0.2	0.222222
15	10	5	2	0.2	0.4	0.266667
16	4	10	1	0.25	0.1	0.142857
17	3	8	2	0.666667	0.25	0.363636
18	6	12	3	0.5	0.25	0.333333
19	6	1	1	0.166667	1	0.285714
20	7	5	2	0.285714	0.4	0.333333
21	5	5	2	0.4	0.4	0.4
22	2	7	2	1	0.285714	0.444444
23	2	4	1	0.5	0.25	0.333333
24	6	11	3	0.5	0.272727	0.352941
25	10	9	4	0.4	0.444444	0.421053
Mean				**0.51**	**0.37**	**0.43**

Fig. 15 Overall performance metrics of the proposed method with different similarity measures for 25 videos from the **SumMe dataset**.

Table 11 Execution time t in seconds taken for evaluation of the user summaries based on the refined matched frames for 50 videos from the **YT dataset**.

	Execution time *(t in sec)*				
User #	1	2	3	4	5
1	1	88	74	71	83
2	89	1	81	68	89
3	70	83	1	58	71
4	68	69	59	1	64
5	78	86	69	64	1

frames selected by various users, the computational overhead will still be feasible which can be seen from Tables 10 and 11 and Fig. 17 for the OV, YT and SM datasets respectively. To evaluate the human consistency of each user summary for 50 videos in the OV and YT dataset, it took on average 52.05 seconds and 74.1 seconds respectively, whereas it took on average 38.6 seconds for

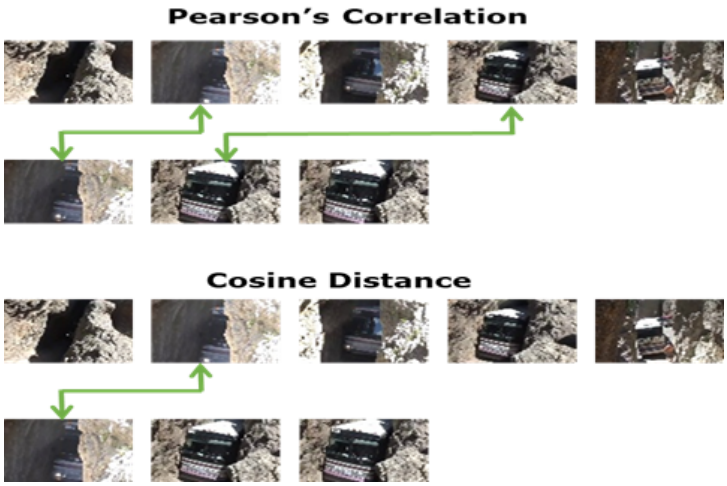

Fig. 16 The found matched frames of the video v5 (Bus in rock tunnel) from the SumMe dataset between the summary of User #1 and overall trimmed summary using Pearsons correlation (top) and cosine similarity (bottom) respectively.

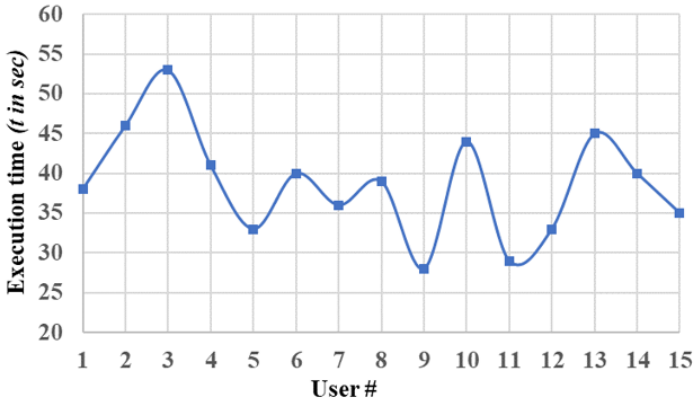

Fig. 17 Execution time t in seconds taken for evaluation of the user summaries based on the refined matched frames for 25 videos from the **SM dataset**.

25 videos in the SM dataset. It revealed that the SM dataset achieves higher computational efficiency due to its less number of videos and the OV dataset performs efficiently compared to the YT dataset due to its short duration.

6 Conclusions and future work

In this paper, we investigated the human consistency evaluation of static video summaries. How to identify the matched frames between the user summaries

plays a crucial role for their performance measurement. The human consistency is investigated from three aspects: potential matched frames, refined matched frames and the number of key frames selected. The potential matched frames are identified through a two-way search without involving any threshold, and the similarity of two frames is measured using the Pearson's correlation coefficient. The potential matched frames may not be reliable, so the consistency between such matched frames is modeled through maintaining the difference between such matched frames, leading the false and weak matches to be eliminated and the correct matched frames to be identified. While the establishment of matched frames involves the challenging task of feature extraction and matching, we further investigate the human consistency using pairwise correlation between the numbers of frames chosen by different users. The experimental results based on several publicly accessible datasets reveal that the users are usually not consistent among themselves. In comparison, the consistency of the user summaries over the OV dataset is better compared to those from the other datasets, due to its quality, predefined structure and similar content. Therefore the following conclusions can be drawn: (i) The datasets selected for performance measurement should be well structured. Otherwise, the technique for video summarization cannot be expected to perform well. The maximum agreement level of the users based on the refined matched frames for any automatic video summaries evaluated using the automatic evaluation method proposed by Kannappan *et al.* [27], for the OV, YT and SM datasets are **0.66**, **0.56** and **0.43** as shown in Figs. 11, 12 and 13 respectively. (ii) The agreement levels among the user summaries may indicate the overall best performance that the automatic techniques can achieve for video summarization. (iii) The agreement level also indicates the complexity of the video contents and if it is low, then it shows that the activities in the video may vary significantly. These results reveal that the performance measures reported by [3], [22] may be too optimistic and misleading.

Even though Gygli*et al.* [24] investigated the consistency of the human summaries and shown the mean f-measures, 0.179, 0.311, ad 0.409, of the worst, mean and best human summaries at 15% summary length for retrieving video skim excerpts from the SM dataset, this is the first detailed, systematic and comprehensive study focusing on the human consistency evaluation of static video summaries over three publicly accessible datasets: OV, YT and SM. It is interesting to note that the human agreement level of 0.43 in mean f-measure over the SM dataset reported in this paper are inline with that of the best human summary reported in [24], while the former has shorter summary length than the latter. We wold extend this work to verify the human consistency of video skims in the near future. Moreover, due to the availability of multi-core CPUs/GPUs [35] and advanced optimization methods [36–38] in scientific computer applications, we will further our work by enhancing the proposed algorithm through utilizing the hardware resources for processing large video data in the near future.

Acknowledgments

The first author would like to thank for the award given by Aberystwyth University under the Departmental Overseas Scholarship (DOS) and partly funding by Object Matrix, Ltd on the project. The authors would express their gratitude to the associate editor and anonymous reviewers for their constructive comments that have improved the readability and quality of this paper.

References

1. Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 1, Oct 1998, pp. 866–870.
2. B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 3, no. 1, p. 3, 2007.
3. S. E. F. De Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.
4. H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia systems*, vol. 1, no. 1, pp. 10–28, 1993.
5. S. Uchihashi and J. Foote, "Summarizing video using a shot importance measure and a frame-packing algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6. IEEE, 1999, pp. 3041–3044.
6. D. Swanberg, C.-F. Shu, and R. C. Jain, "Knowledge-guided parsing in video databases," in *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1993, pp. 13–24.
7. M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer vision and image understanding*, vol. 71, no. 1, pp. 94–109, 1998.
8. A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580–588, 1999.
9. O. Javed, Z. Rasheed, and M. Shah, "A framework for segmentation of talk and game shows," in *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. 2. IEEE, 2001, pp. 532–537.
10. C. Wang, H. Yang, and C. Meinel, "Exploring multimodal video representation for action recognition," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 1924–1931.
11. C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 988–997.
12. C. Wang, H. Yang, and C. Meinel, "Image captioning with deep bidirectional lstms and multi-task learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, p. 40, 2018.
13. B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2017.
14. C. Wang, H. Yang, and C. Meinel, "Deep semantic mapping for cross-modal retrieval," in *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*. IEEE, 2015, pp. 234–241.
15. ——, "A deep semantic framework for multimodal representation learning," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9255–9276, 2016.

16. K. Li, F.-Z. He, and H.-P. Yu, "Robust visual tracking based on convolutional features with illumination and occlusion handing," *Journal of Computer Science and Technology*, vol. 33, no. 1, pp. 223–236, 2018.
17. K. Li, F.-z. He, H.-p. Yu, and X. Chen, "A correlative classifiers approach based on particle filter and sample set for tracking occluded target," *Applied Mathematics-A Journal of Chinese Universities*, vol. 32, no. 3, pp. 294–312, 2017.
18. X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," in *Proceedings of the 10th International Multimedia Modelling Conference.* IEEE, 2004, pp. 117–123.
19. T. Liu, X. Zhang, J. Feng, and K.-T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern recognition letters*, vol. 25, no. 12, pp. 1451–1457, 2004.
20. M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "Stimo: Still and moving video storyboard for the web scenario," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 47–69, 2010.
21. S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.
22. K. M. Mahmoud, "An enhanced method for evaluating automatic video summaries," *arXiv preprint arXiv:1401.3590*, 2014.
23. K. Mahmoud, N. Ghanem, and M. Ismail, "Vgraph: an effective approach for generating static video summaries," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 811–818.
24. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision (ECCV).* Springer, 2014, pp. 505–520.
25. A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2127–2136.
26. S. Kannappan, Y. Liu, and B. P. Tiddeman, "Performance evaluation of video summaries using efficient image euclidean distance," in *International Symposium on Visual Computing (ISVC).* Springer, 2016, pp. 33–42.
27. S. Kannappan, Y. Liu, and B. Tiddeman, "A pertinent evaluation of automatic video summary," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR).* IEEE, 2016, pp. 2240–2245.
28. K. Yen, E. K. Yen, and R. G. Johnston, "The ineffectiveness of the correlation coefficient for image comparisons," 1996.
29. A. Egozi, Y. Keller, and H. Guterman, "Improving shape retrieval by spectral matching and meta similarity," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1319–1327, 2010.
30. L. Lovstakken, S. Bjaerum, K. Kristoffersen, R. Haaverstad, and H. Torp, "Real-time adaptive clutter rejection filtering in color flow imaging using power method iterations," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 53, no. 9, pp. 1597–1608, Sept 2006.
31. H. M. Blanken, H. E. Blok, L. Feng, and A. P. Vries, *Multimedia retrieval.* Springer, 2007.
32. Open video project. [Online]. Available: http://www.open-video.org
33. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European Conference on Computer Vision (ECCV).* Springer, 2016, pp. 766–782.
34. P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.
35. Y. Zhou, F. He, N. Hou, and Y. Qiu, "Parallel ant colony optimization on multi-core simd cpus," *Future Generation Computer Systems*, vol. 79, pp. 473–487, 2018.
36. X. Yan, F. He, N. Hou, and H. Ai, "An efficient particle swarm optimization for large-scale hardware/software co-design system," *International Journal of Cooperative Information Systems (IJCIS)*, vol. 27, no. 01, p. 1741001, 2018.
37. Y. Zhou, F. He, and Y. Qiu, "Dynamic strategy based parallel ant colony optimization on gpus for tsps," *Science China Information Sciences*, vol. 60, no. 6, p. 068102, 2017.

38. X.-H. Yan, F.-Z. He, and Y.-L. Chen, "A novel hardware/software partitioning method based on position disturbed particle swarm optimization with invasive weed optimization," *Journal of Computer Science and Technology (JCST)*, vol. 32, no. 2, pp. 340–355, 2017.